Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models

Zitong Yang, Yu Bai, Song Mei

April 13, 2021

Reported by Chendi Wang in the deep learning seminar at CUHK-SZ

Outline

- Problem Formulation
- Assumptions
- Main Theorem
- (Inferred) Asymptotic Power Laws
- Sketch of Proofs

Model setup

- Consider a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with n samples.
- $\mathbf{x}_i \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $y_i = f_d(\mathbf{x}_i) + \epsilon_i$, where the noises $\epsilon_i \sim_{i.i.d.} \mathcal{N}(0, \tau^2)$ with $\tau^2 \geq 0$ are independent of $\{\mathbf{x}_i\}_{i=1}^n$.
- Let $(\boldsymbol{\theta}_j)_{j=1}^N \sim_{i.i.d.} \mathrm{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$.
- Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, define the random features function class $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ by

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta}) = \left\{ f(\mathbf{x}) = \sum_{j=1}^{N} a_j \sigma(\langle \mathbf{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d}) : \boldsymbol{a} \in \mathbb{R}^N \right\}.$$

Generalization error and minimum norm interpolator

- Population risk: $R(\boldsymbol{a}) = \mathbb{E}_{\mathbf{x},y}\left(y - \sum_{j=1}^{N} a_j \sigma\left(\left\langle \mathbf{x}, \boldsymbol{\theta}_j \right\rangle / \sqrt{d}\right)\right)^2$.
- Empirical risk: $\hat{R}_n(\boldsymbol{a}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{N} a_j \sigma\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}_j \right\rangle / \sqrt{d}\right)\right)^2$.
- Denote the regularized empirical risk minimizer with vanishing regularization by

$$\boldsymbol{a}_{\min} = \lim_{\lambda \to 0} \arg\min_{\boldsymbol{a}} \left[\hat{R}_n(\boldsymbol{a}) + \lambda \left\|\boldsymbol{a}\right\|_2^2\right].$$

- Note that $\hat{R}_n(\boldsymbol{a})$ is quadratic for random feature models and $\boldsymbol{a}_{\min}$ can be interpreted as the minimum ℓ_2 norm interpolator if $\min_{\boldsymbol{a}} \hat{R}_n(\boldsymbol{a}) = \hat{R}_n(\boldsymbol{a}_{\min}) = 0$, that is, $\boldsymbol{a}_{\min}$ is the solution to

$$\min_{\boldsymbol{a}} \left\|\boldsymbol{a}\right\|_2 \text{ s.t. } \hat{R}_n(\boldsymbol{a}) = 0$$

- Generalization error: $R(N, n, d) = R(\boldsymbol{a}_{\min})$.

Uniform convergence bounds

- Uniform convergence bound over a norm ball:

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \left(R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a}) \right).$$

- Uniform convergence over interpolators in the norm ball:

$$T(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \hat{R}_n(\boldsymbol{a}) = 0} R(\boldsymbol{a}).$$

- We need $\hat{R}_n(\boldsymbol{a}_{\min}) = 0$ and take $A \ge (N/d)\|\boldsymbol{a}_{\min}\|_2^2$ to have a non-empty feasible region.
- For $A \ge (N/d)\|\boldsymbol{a}_{\min}\|$, there holds

$$U(A, N, n, d) \ge T(A, N, n, d) \ge R(\boldsymbol{a}_{\min}).$$

Assumptions

- **Assumption 1** (Linear target function). $f_d \in L^2\left(\mathbb{S}^{d-1}\left(\sqrt{d}\right)\right)$ with $f_d(\mathbf{x}) = \left\langle \boldsymbol{\beta}^{(d)}, \mathbf{x} \right\rangle$, where $\boldsymbol{\beta}^{(d)} \in \mathbb{R}^d$ and $\lim_{d\to\infty} \|\boldsymbol{\beta}^{(d)}\|_2^2 = F_1^2$.
- **Assumption 2** (Activation function). Let $\sigma \in C^2(\mathbb{R})$ with $|\sigma(u)|, |\sigma'(u)|, |\sigma''(u)| \leq c_0 e^{c_1|u|}$ for some constant $c_0, c_1 < \infty$. Define

$$\mu_0 = \mathbb{E}\left[\sigma(G)\right], \mu_1 = \mathbb{E}\left[G\sigma(G)\right], \mu_*^2 = \mathbb{E}\left[\sigma(G)^2\right] - \mu_0^2 - \mu_1^2,$$

  where the expectation is w.r.t. $G \sim \mathcal{N}(0,1)$. Assume $\mu_0 = 0$, $0 < \mu_1^2, \mu_*^2 < \infty$.
- **Assumption 3** (Proportional limit). Let $N = N(d)$ and $n = n(d)$. Assume that the following limits exist in $(0, \infty)$:

$$\lim_{d\to\infty} N/d = \psi_1, \lim_{d\to\infty} n/d = \psi_2.$$

- Other technical assumptions used in the proof.

Main Theorem

Theorem

Under Assumption 1, Assumption 2, Assumption 3, and other technical assumptions, there hold the following conclusions.

1. *For any* $A \in \Gamma_U$, *we have*

$$U(A, N, n, d) = \mathcal{U}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).$$

2. *For any* $A \in \Gamma_T$, *we have*

$$T(A, N, n, d) = \mathcal{T}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).$$

- Here the sets Γ_U and Γ_T will be defined later.
- The quantities $\mathcal{U}$ and $\mathcal{T}$ are involved and will be defined later.
- In a special case in this paper, the inferred asymptotic power law of $\mathcal{U}$ and $\mathcal{T}$ is given.

High dimensional regime

- In this paper, they consider the case where $d \to +\infty$.
- Denote $\hat{\psi}_1 = N/d$ and $\hat{\psi}_2 = n/d$. Recall the constrained $(N/d)\|\boldsymbol{a}\|_2^2 \leq A$ and the generalization error $R(\boldsymbol{a}_{\min})$.
- In addition to $\mathcal{U}$ and $\mathcal{T}$, Theorem 1 of Mei & Montanari (2019) implies the following convergence (in probability)

$$\hat{\psi}_1 \|\boldsymbol{a}_{\min}\|_2^2 \overset{d \to +\infty}{\longrightarrow} \mathcal{A}(\psi_1, \psi_2),$$
$$R(\boldsymbol{a}_{\min}) \overset{d \to +\infty}{\longrightarrow} \mathcal{R}(\psi_1, \psi_2).$$

- Here $\mathcal{A}$ and $\mathcal{R}$ are defined in Mei & Montanari (2019).

Kernel regime

- As $N \to \infty$, the random feature space $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ (equipped with proper inner product) converges to an RKHS (reproducing kernel Hilbert space) induce by the kernel

$$H(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \left[\sigma\left(\langle \mathbf{x}, \boldsymbol{\theta} \rangle\right) \sigma\left(\langle \mathbf{x}', \boldsymbol{\theta} \rangle\right) \right].$$

- They expect that if they take $\psi_1 \to +\infty$ after $N, d, n \to +\infty$, the formula of $\mathcal{U}$ and $\mathcal{T}$ will coincide with the corresponding asymptotic limit of U and T for kernel ridge regression with the kernel H. (? an intuition)
- Denote

$$\mathcal{U}_\infty(A, \psi_2) = \lim_{\psi_1 \to \infty} \mathcal{U}(A, \psi_1, \psi_2), \quad \mathcal{T}_\infty(A, \psi_2) = \lim_{\psi_1 \to \infty} \mathcal{T}(A, \psi_1, \psi_2),$$
$$\mathcal{A}_\infty(\psi_2) = \lim_{\psi_1 \to +\infty} \mathcal{A}(\psi_1, \psi_2), \quad \mathcal{R}_\infty(\psi_2) = \lim_{\psi_1 \to \infty} \mathcal{R}(\psi_1, \psi_2).$$

Low norm uniform convergence bounds

- How to choose norm A in $\mathcal{U}$ and $\mathcal{T}$?
- We need at least $A \geq \hat{\psi}_1 \|\boldsymbol{a}_{\min}\|_2^2$. Therefore, we will choose

$$A = \alpha \hat{\psi}_1 \|\boldsymbol{a}_{\min}\|_2^2, \quad \text{for some } \alpha > 1.$$

- Note that $\hat{\psi}_1 \|\boldsymbol{a}_{\min}\|_2^2 \to \mathcal{A}(\psi_1, \psi_2)$ as $d \to +\infty$. For a fixed $\alpha > 1$, we further define

$$\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) = \mathcal{U}(\alpha \mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \quad \mathcal{U}_\infty^{(\alpha)} = \lim_{\psi_1 \to \infty} \mathcal{U}^{(\alpha)}(\psi_1, \psi_2),$$

and

$$\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) = \mathcal{T}(\alpha \mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \quad \mathcal{T}_\infty^{(\alpha)} = \lim_{\psi_1 \to \infty} \mathcal{T}^{(\alpha)}(\psi_1, \psi_2).$$

Inferred asymptotic power law, I

- Norm of the minimum norm interpolator:

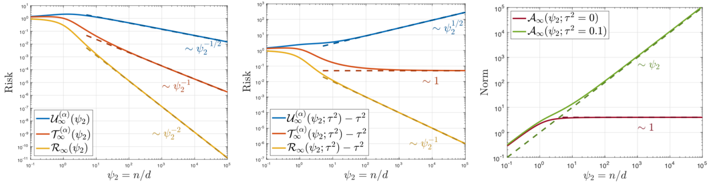
$$\mathcal{A}_\infty(\psi_2; \tau^2 > 0) \sim \psi_2, \quad \mathcal{A}_\infty(\psi_2; \tau^2 = 0) \sim 1.$$

- Kernel regime with noiseless data ($\tau^2 = 0$):

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2) \sim \psi_2^{-1/2}, \quad \mathcal{T}_\infty^{(\alpha)} \sim \psi_2^{-1}, \quad \mathcal{R}_\infty(\psi_2) \sim \psi_2^{-2}.$$

- Kernel regime with noiseless data ($\tau^2 > 0$):

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2) - \tau^2 \sim \psi_2^{1/2}, \quad \mathcal{T}_\infty^{(\alpha)} - \tau^2 \sim 1, \quad \mathcal{R}_\infty(\psi_2) - \tau^2 \sim \psi_2^{-1}.$$



(a) Noiseless response ($\tau^2 = 0$) (b) Noisy response ($\tau^2 = 0.1$) (c) Minimum norm $\mathcal{A}_\infty(\psi_2)$

Figure 1. Random feature regression with activation function $\sigma(x) = \max(0, x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and $\psi_1 = \infty$. The horizontal axes are the number of samples $\psi_2 = \lim_{d\to\infty} n/d$. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_2^p in the log scale. Figure 1(a) and 1(b): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (17), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (18), red curve), the risk of minimum norm interpolator (Eq. (13), yellow curve). Figure 1(c): Minimum norm required to interpolate the training data (Eq. (12)).

Inferred asymptotic power law, II

- The divergence of $\mathcal{U}_\infty^{(\alpha)}$ with noisy data is partly due to that $\mathcal{A}_\infty(\psi_2)$ blows up linearly in ψ_2.
- In fact, they can develop a heuristic intuition that $\mathcal{U}_\infty(A, \psi_2) \sim A/\psi_2^{1/2}$.

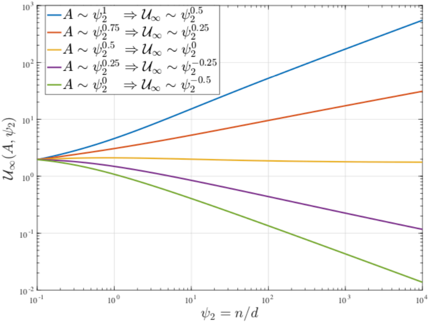


Figure 3. Uniform convergence $\mathcal{U}_\infty(A(\psi_2), \psi_2)$ over the norm ball in the kernel regime $\psi_1 \to \infty$. The size of the norm ball $A = A(\psi_2)$ is chosen according to different power laws as shown in the legend.

Inferred asymptotic power law, III

- Finite-width regime:

$$\mathcal{U}^{(\alpha)}(\psi_1,\psi_2) - \mathcal{U}_\infty^{(\alpha)}(\psi_2) \sim \psi_1^{-1},$$
$$\mathcal{T}^{(\alpha)}(\psi_1,\psi_2) - \mathcal{T}_\infty^{(\alpha)}(\psi_2) \sim \psi_1^{-1},$$
$$\mathcal{R}(\psi_1,\psi_2) - \mathcal{R}_\infty(\psi_2) \sim \psi_1^{-1},$$
$$\mathcal{A}(\psi_1,\psi_2) - \mathcal{A}_\infty(\psi_2) \sim \psi_1^{-1}.$$

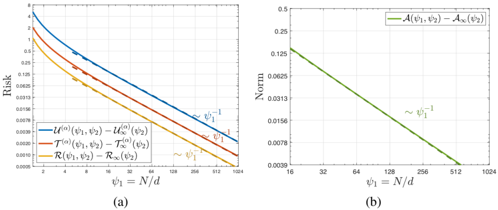


Figure 4. Random feature regression with the number of sample $\psi_2 = 1.5$, activation function $\sigma(x) = \max(0,x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and noise level $\tau^2 = 0.1$. The horizontal axes are the number of features ψ_1. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_1^p in the log scale. Figure 4(a): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (15), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (16), red curve), the risk of minimum norm interpolator (Eq. (9), yellow curve). Figure 4(b): Minimum norm required to interpolate the training data (Eq. (8)).

Some notations

- Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$, $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N)^\top \in \mathbb{R}^{N \times d}$, and $\mathbf{y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$.
- Denote $\mathbf{v} = (v_i)_{i \in [n]} \in \mathbb{R}^n$, $\mathbf{U} = (U_{ij})_{i,j \in [N]} \in \mathbb{R}^{N \times N}$, and $\mathbf{Z} = (Z_{ij})_{i \in [n], j \in [N]} \in \mathbb{R}^{n \times N}$ with

$$\begin{aligned} v_i &= \mathbb{E}_{\epsilon, \mathbf{x}} \left[y \sigma \left(\langle \mathbf{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d} \right) \right], \\ U_{ij} &= \mathbb{E}_{\mathbf{x}} \left[\sigma \left(\langle \mathbf{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d} \right) \sigma \left(\langle \mathbf{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d} \right) \right], \\ Z_{ij} &= \sigma \left(\langle \mathbf{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d} \right) / \sqrt{d}. \end{aligned}$$

- Rewrite

$$\begin{aligned} R(\boldsymbol{a}) &= \langle \boldsymbol{a}, \mathbf{U} \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \mathbf{v} \rangle + \mathbb{E}[y^2], \\ \hat{R}_n(\boldsymbol{a}) &= \frac{1}{n} \left\| \mathbf{y} - \sqrt{d} \mathbf{Z} \boldsymbol{a} \right\|_2^2 \\ &= \hat{\psi}_2^{-1} \langle \boldsymbol{a}, \mathbf{Z}^\top \mathbf{Z} \boldsymbol{a} \rangle - 2 \hat{\psi}_2^{-1} \frac{\langle \mathbf{Z}^\top \mathbf{y}, \boldsymbol{a} \rangle}{\sqrt{d}} + \frac{1}{n} \|\mathbf{y}\|_2^2. \end{aligned}$$

Strong duality

- Recall

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \left(R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a})\right),$$

$$T(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \hat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}).$$

- Let

$$\overline{U}(\lambda, N, n, d) = \sup_{\boldsymbol{a}} \left[R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a}) - \hat{\psi}_1 \lambda \|\boldsymbol{a}\|_2^2\right],$$

$$\overline{T}(\lambda, N, n, d) = \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \left[R(\boldsymbol{a}) - \hat{\psi}_1 \lambda \|\boldsymbol{a}\|_2^2 + 2\left\langle \boldsymbol{\mu}, \mathbf{Z}\boldsymbol{a} - \mathbf{y}/\sqrt{d}\right\rangle\right].$$

Proposition 1

For any $A > 0$, there holds

$$U(A, N, n, d) = \inf_{\lambda \ge 0} \left[\overline{U}(\lambda, N, n, d) + \lambda A\right].$$

Moreover, for any $A > \hat{\psi}_1 \|\boldsymbol{a}_{\min}\|_2^2$, there holds

$$T(A, N, n, d) = \inf_{\lambda \ge 0} \left[\overline{T}(\lambda, N, n, d) + \lambda A\right].$$

Limit of dual forms

Proposition 2

Let assumptions in the main theorem hold. Then for $\lambda \in \Lambda_U$, with high probability the maximizer in the definition of $\overline{U}$ can be achieved at a unique point $\overline{\boldsymbol{a}}_U(\lambda)$ and we have

$$\begin{aligned}\overline{U}(\lambda, N, n, d) &= \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \\ \hat{\psi}_1 \|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 &= \mathcal{A}_U(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).\end{aligned}$$

Moreover, for $\lambda \in \Lambda_T$, with high probability the maximizer in the definition of $\overline{T}$ can be achieved at a unique point $\overline{\boldsymbol{a}}_T(\lambda)$ and we have

$$\begin{aligned}\overline{T}(\lambda, N, n, d) &= \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \\ \hat{\psi}_1 \|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2 &= \mathcal{A}_T(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).\end{aligned}$$

- The sets Λ_U and Λ_T will be given later in the proof.
- The definitions of $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$ are given in the appendix.

Heuristic formulae of quantities in Proposition 2

Remark 1. *Here we present the heuristic formulae of $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$, and defer their rigorous definition to the appendix. Define a function $g_0(\boldsymbol{q};\boldsymbol{\psi})$ by*

$$
\begin{aligned}
g_0(\boldsymbol{q};\boldsymbol{\psi}) \equiv\ & \mathrm{ext}_{z_1,z_2}\Big[\log\big((s_2z_1+1)(t_2z_2+1)\\
& -\mu_1^2(1+p)^2z_1z_2\big) - \mu_\star^2 z_1z_2 + s_1z_1 + t_1z_2 \\
& -\psi_1\log(z_1/\psi_1) - \psi_2\log(z_2/\psi_2) - \psi_1 - \psi_2\Big],
\end{aligned} \tag{22}
$$

where ext *stands for setting z_1 and z_2 to be stationery (which is a common symbol in statistical physics heuristics). We then take*

$$
\overline{\mathcal{U}}(\lambda,\boldsymbol{\psi}) = F_1^2(1-\mu_1^2\gamma_{s_2} - \gamma_p - \gamma_{t_2}) + \tau^2(1-\gamma_{t_1}),
$$

where $\gamma_a \equiv \partial_a g_0(\boldsymbol{q};\boldsymbol{\psi})|_{\boldsymbol{q}=(\mu_\star^2-\lambda\psi_1,\mu_1^2,\psi_2,0,0)}$ for the symbol $a \in \{s_1,s_2,t_1,t_2,p\}$, and

$$
\overline{\mathcal{T}}(\lambda,\boldsymbol{\psi}) = F_1^2(1-\mu_1^2\nu_{s_2} - \nu_p - \nu_{t_2}) + \tau^2(1-\nu_{t_1}),
$$

where we define $\nu_a \equiv \partial_a g_0(\boldsymbol{q};\boldsymbol{\psi})|_{\boldsymbol{q}=(\mu_\star^2-\lambda\psi_1,\mu_1^2,0,0,0)}$ for symbols $a \in \{s_1,s_2,t_1,t_2,p\}$. Finally $\mathcal{A}_U = -\partial_\lambda\overline{\mathcal{U}}$, $\mathcal{A}_T = -\partial_\lambda\overline{\mathcal{T}}$. By a further simplification, we can express these formulae to be rational functions of $(\mu_1^2,\mu_\star^2,\lambda,\psi_1,\psi_2,m_1,m_2)$ where (m_1,m_2) is the stationery point of the variational problem in Eq. (22) (c.f. Remark 2).

- Here $\mathbf{q} = (s_1,s_2,t_1,t_2,p)$ and $\boldsymbol{\psi} = (\psi_1,\psi_2)$.

Formulae for uniform convergence bounds

- For $A \in \Gamma_U = \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, define

$$\mathcal{U}(A, \psi_1, \psi_2) = \inf_{\lambda \geq 0} \left[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A \right].$$

- For $A \in \Gamma_T = \{\mathcal{A}_T(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_T\}$, define

$$\mathcal{T}(A, \psi_1, \psi_2) = \inf_{\lambda \geq 0} \left[\overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + \lambda A \right].$$

Key point of the proof of Proposition 1

- Strong duality holds for quadratic program with single quadratic constraint.
- For U, there holds

$$\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \left(R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a}) \right) = \inf_{\lambda \ge 0} \sup_{\boldsymbol{a}} \left[R(\boldsymbol{a}) - \hat{R}_n(\boldsymbol{a}) - \lambda \left(\hat{\psi}_1 \|\boldsymbol{a}\|_2^2 - A \right) \right].$$

- $\{\boldsymbol{a} : \hat{R}_n(\boldsymbol{a}) = 0\} = \{\boldsymbol{a}_{\min} + \mathbf{R}\mathbf{u}, \mathbf{u} \in \mathbb{R}^m\}$, where $m = \dim(\mathrm{Null}(\mathbf{Z}))$ and $\mathbf{R} \in \mathbb{R}^{N \times m}$ is a matrix such that $\mathrm{Span}(\mathbf{R}) = \mathrm{Null}(\mathbf{Z})$.
- There holds

$$\begin{aligned}
&\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A,\, \hat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}) \\
=& R(\boldsymbol{a}_{\min}) + \sup_{\|\mathbf{R}\mathbf{u} + \boldsymbol{a}_{\min}\|_2^2 \le \hat{\psi}_1^{-1} A} \left[\langle \mathbf{u}, \mathbf{R}^\top \mathbf{U} \mathbf{R} \mathbf{u} \rangle + 2 \langle \mathbf{R}\mathbf{u}, \mathbf{U}\boldsymbol{a}_{\min} - \mathbf{v} \rangle \right] \\
=& \inf_{\lambda \ge 0} \left\{ \lambda A + \sup_{\hat{R}_n(\boldsymbol{a})=0} \left[R(\boldsymbol{a}) - \lambda \hat{\psi}_1 \|\boldsymbol{a}\|_2^2 \right] \right\}
\end{aligned}$$

Proof sketch of Proposition 2

- The definitions of $\overline{U}$ and $\overline{T}$ depend on $\boldsymbol{\beta} = \boldsymbol{\beta}^{(d)}$ such that $f_d(\mathbf{x}) = \left\langle \boldsymbol{\beta}^{(d)}, \mathbf{x} \right\rangle$.
- Since $\mathbf{x}_i's$ ans $\boldsymbol{\theta}_i's$ are rotationally invariant, there holds $\overline{U}(\boldsymbol{\beta}_1, \lambda, N, n, d)$ $= \overline{U}(\boldsymbol{\beta}_2, \lambda, N, n, d)$ and $\overline{T}(\boldsymbol{\beta}_1, \lambda, N, n, d) = \overline{T}(\boldsymbol{\beta}_2, \lambda, N, n, d)$ for $\|\boldsymbol{\beta}_1\|_2 = \|\boldsymbol{\beta}_2\|_2$.
- In the proof, they work with the assumption that $\boldsymbol{\beta}^{(d)} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$.

Proof sketch of Proposition 2

- Recall the matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$ with $U_{ij} = \mathbb{E}_{\mathbf{x}}\left[\sigma\left(\langle \mathbf{x}, \boldsymbol{\theta}_i\rangle / \sqrt{d}\right)\sigma\left(\langle \mathbf{x}, \boldsymbol{\theta}_j\rangle / \sqrt{d}\right)\right].$
- Let $\mathbf{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^\top / d$ and let μ_1, μ_* be defined in Assumption 2.
- There holds the following decomposition

$$\mathbf{U} = \mu_1^2 \mathbf{Q} + \mu_*^2 \mathbf{I}_N + \boldsymbol{\Delta}$$

  with $\boldsymbol{\Delta}$ being a perturbation such that $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$.
- This decomposition was first proved by El Karoui (2010) for the Gaussian case and has been widely used in studying the interpolation regime (c.f., Liang & Rakhlin, 2020).
- Nonlinear $\to$ linear in high dimensions ($d \to +\infty$ (e.g.,this paper) or sufficiently large d (e.g., Liang & Rakhlin, 2020)).
- This decomposition requires the smoothness of σ.

Proof sketch of Proposition 2

In the following, we would like to show that $\boldsymbol{\Delta}$ has vanishing effects in the asymptotics of $\overline{U}$, $\overline{T}$, $\|\overline{\boldsymbol{a}}_U\|_2^2$ and $\|\overline{\boldsymbol{a}}_T\|_2^2$. For this purpose, we denote

$$\begin{aligned}
\boldsymbol{U}_c &= \mu_1^2 \boldsymbol{Q} + \mu_\star^2 \mathbf{I}_N, \\
R_c(\boldsymbol{a}) &= \langle \boldsymbol{a}, \boldsymbol{U}_c \boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \boldsymbol{v} \rangle + \mathbb{E}[y^2], \\
\widehat{R}_{c,n}(\boldsymbol{a}) &= \langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} \boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{y}/\sqrt{d} \rangle + \mathbb{E}[y^2], \\
\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(R_c(\boldsymbol{a}) - \widehat{R}_{c,n}(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 \Big), \\
\overline{T}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \Big[R_c(\boldsymbol{a}) - \lambda \psi_1 \|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d} \,\rangle \Big].
\end{aligned}$$

- In the notations defined above, ψ_1 and ψ_2 should be $\hat{\psi}_1$ and $\hat{\psi}_2$, respectively.
- There holds

$$\overline{U}_c(\lambda, N, n, d) = \sup_{\boldsymbol{a}} \left(\left\langle \boldsymbol{a}, \overline{\mathbf{M}} \boldsymbol{a} \right\rangle - 2 \left\langle \boldsymbol{a}, \overline{\mathbf{v}} \right\rangle \right)$$

with $\overline{\mathbf{M}} = \mathbf{U}_c - \hat{\psi}_2^{-1} \mathbf{Z}^\top \mathbf{Z} - \hat{\psi}_1 \lambda \mathbf{I}_N$ and $\overline{\mathbf{v}} = \mathbf{v} - \hat{\psi}_2^{-1} \mathbf{Z}^\top \mathbf{y}/\sqrt{d}$.

Proof sketch of Proposition 2

- Assume that there exists $\delta > 0$ and $\lambda_U = \lambda_U(\psi_1, \psi_2, \mu_1^2, \mu_*^2)$ such that for any fixed $\lambda \in \Lambda_U = (\lambda_U, +\infty)$, there holds

$$\overline{\mathbf{M}} = \overline{\mathbf{M}}(\lambda) \preccurlyeq -\delta \mathbf{I}_N.$$

- For $\lambda \in \Lambda_U$, there holds $\overline{\boldsymbol{a}}_{U,c}(\lambda) = \overline{\mathbf{M}}^{-1} \overline{\mathbf{v}}$.
- Note that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\|\boldsymbol{\Delta}\|_{\mathrm{op}} \leq \delta/2$ with high probability for d large enough.
- $\overline{\boldsymbol{a}}_U(\lambda) = \left(\overline{\mathbf{M}} + \boldsymbol{\Delta}\right)^{-1} \overline{\mathbf{v}}$ for $\lambda \in \Lambda_U$ and d large enough.
- They have

$$\begin{aligned} \|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 &= (1 + o_{d,\mathbb{P}}) \left\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\right\|_2^2, \\ \overline{U}_c(\lambda, N, n, d) &= \overline{U}(\lambda, N, n, d) + o_{d,\mathbb{P}} \left(\left\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\right\|_2^2 + 1 \right). \end{aligned}$$

Proof sketch of Proposition 2

By Eq. (37) and (38), simple calculation shows that

$$\overline{U}_c(\lambda, N, n, d) \equiv -\langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}} \rangle = -\Psi_1 - \Psi_2 - \Psi_3,$$
$$\|\overline{\boldsymbol{a}}_{U,c}\|_2^2 \equiv \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-2} \overline{\boldsymbol{v}} \rangle = \Phi_1 + \Phi_2 + \Phi_3,$$

where

$$\Psi_1 = \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, \qquad \Phi_1 = \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle,$$
$$\Psi_2 = -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, \qquad \Phi_2 = -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle,$$
$$\Psi_3 = \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}} \rangle, \qquad \Phi_3 = \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \frac{\boldsymbol{Z}^\top \boldsymbol{y}}{\sqrt{d}} \rangle.$$

- Here ψ_1 (ψ_2) should be $\hat{\psi}_1$ ($\hat{\psi}_2$).

Proof sketch of Proposition 2

Proposition 5. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, denote $\boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, then we have*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] \xrightarrow{\mathbb{P}} \mu_1^2 F_1^2 \cdot \partial_{s_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] \xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] \xrightarrow{\mathbb{P}} F_1^2 \cdot \Big(\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big) + \tau^2 \Big(\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] \xrightarrow{\mathbb{P}} -\mu_1^2 F_1^2 \cdot \partial_{s1}\partial_{s2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] \xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s1}\partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] \xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s_1}\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),
\end{aligned}$$

where $\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi})$ for $k \in \{1, 2\}$ stands for the k'th derivatives (as a vector or a matrix) of $g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi}) = \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)] \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \qquad \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

where the definitions of $\overline{\mathcal{U}}$ and $\mathcal{A}_U$ are given in Definition 5. Here $\xrightarrow{\mathbb{P}}$ stands for convergence in probability as $N/d \to \psi_1$ and $n/d \to \psi_2$ (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$).

- The idea of the proof of Proposition 5 follows mainly (Mei & Montanari, 2019)

Proof sketch of Proposition 2

Lemma 2. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, we have*

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] = o_{d,\mathbb{P}}(1),$$
$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] = o_{d,\mathbb{P}}(1),$$

so that

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] = o_{d,\mathbb{P}}(1).$$

Here, $o_{d,\mathbb{P}}(1)$ stands for converges to 0 in probability (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$) as $N/d \to \psi_1$ and $n/d \to \psi_2$ and $d \to \infty$.

Now, combining Lemma 2 and Proposition 5, we have

$$\overline{U}_c(\lambda, N, n, d) \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \quad \psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2 \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

Proof sketch of Proposition 2

- The proof of the results of T is similar to the proof of U by replacing $\overline{\mathbf{M}}$ and $\overline{\mathbf{v}}$ with $\widetilde{\mathbf{M}}$ and $\tilde{\mathbf{v}}$, accordingly. Here

$$\widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{U}_c - \hat{\psi}_1 \lambda \mathbf{I}_N & \mathbf{Z}^\top \\ \mathbf{Z} & \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{v}} = \begin{bmatrix} \mathbf{v} \\ \mathbf{y}/\sqrt{d} \end{bmatrix}.$$

- Let $\mathbf{P}_{\mathrm{Null}} = \mathbf{I}_N - \mathbf{Z}^\dagger \mathbf{Z}$ be the projection onto $\mathrm{Null}(\mathbf{Z})$. Assume that there exists $\delta > 0$ and $\lambda_T = \lambda_T(\psi_1, \psi_2, \mu_1^2, \mu_*^2)$ such that for any fixed $\lambda \in \Lambda_T = (\lambda_T, \infty)$, there holds

$$\mathbf{P}_{\mathrm{Null}} \left[\mu_1^2 \mathbf{Q} + (\mu_*^2 - \hat{\psi}_1 \lambda) \mathbf{I}_N \right] \mathbf{P}_{\mathrm{Null}} \preccurlyeq -\delta \mathbf{P}_{\mathrm{Null}},$$

  and $\mathbf{Z}$ has full rank with $\sigma_{\min}(\mathbf{Z}) \geq \delta$.
- $\widetilde{\mathbf{M}} = \widetilde{\mathbf{M}}(\lambda)$ is invertible for $\lambda \in \Lambda_T$.

Proof sketch of the main theorem

- For $A \in \Gamma_U = \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, we have

  $$\lambda_* = \lambda_*(A) = \inf_{\lambda} \{\lambda : \mathcal{A}_U(\lambda, \psi_1, \psi_2) = A\} \in \operatorname{Arg}\min_{\lambda \geq 0} \left[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A\right].$$

  (easy to see?)
- $\overline{\mathcal{U}}(\lambda_*, \psi_1, \psi_2) + \lambda_* A = \mathcal{U}(A, \psi_1, \psi_2)$.
- $U(A, N, n, d) \leq \overline{U}(\lambda_*, N, n, d) + \lambda_* A$ (primal $\leq$ dual for max problem)
- $U(A + \delta, N, n, d) \geq \overline{U}(\lambda_*, N, n, d) + \lambda_*(A - \delta)$ for any $\delta > 0$ with high probability.

Thank You!